SUPERVISED LEARNING REGRESSION

TOTAL MARKS:70

DATA DESCRIPTION: The data set consists of complete educational details of students right from their schooling to MBA and previous work experience. Our main objective is to predict the Salary of the students based on the info available. The data set consists of 391 observations and 19 variables.

ATTRIBUTES:

SlNo	-	ID of the student
Gender	-	Gender of Student
Percent_SSC	-	Percentage of marks scored in SSC
Board_SSC	-	Types of Boards in SSC
Percent_HSC	-	Percentage of marks scored in HSC
Board_HSC	-	Types of Boards in HSC
Stream_HSC	-	Specialization in HSC
Percent Degree	-	Percentage of marks scored in Degree
Course_Degree	-	Different courses in degree
Experience_Yrs	-	Work Experience of the Students
Entrance_Test	-	Test which students give for MBA college Entrance
Percentile_ET	-	Percentage of marks scored in Entrance_Test
Percent_MBA	-	Percentage of marks scored in MBA
Specialization_MBA	-	Specialization in MBA
Marks Communication	-	Percentage of marks scored in Communication
Marks_Projectwork	-	Percentage of marks scored in Project Work
Marks_BOCA activities	-	Percentage of marks scored by students in Extra Curricular
Placement	-	Whether Student got placed or not
Salary	-	Salary of students

1. **Read the dataset (tab, csv, xlsx, txt, inbuilt dataset)**
2. **Summarize important observations from the data set (5 MARKS)**
   Some pointers which would help you, but don't be limited by these
   a. What are the number of rows and no. of cols & types of variables.
   b. Calculate statistical summary for numerical variables
   c. Do an appropriate plot and tell which category of bedroom are the most commonly found in the city according to our data.
3. **Summarize relationships among variables (5 MARKS)**
   a. Plot correlation plots. Which are the variables most correlated with Target? Do you want to exclude some variables from the model based on this analysis? What other actions will you take?
4. **Check for defects in the data. Perform necessary actions to 'fix' these defects (10 MARKS)**

a. Is the target variable normally distributed? If not, rectify it.
b. Do variables have missing/null values?
c. Does the data have outliers?
d. Based on your observations what transformation of features or creation of additional features would you do?

5. **Split dataset into train and test (70:30) (5 MARKS)**
   a. Are both train and test representative of the overall data? How would you ascertain this statistically?
6. **Fit a base model. Please write your key observations (15 MARKS)**
   a. What is the overall R2? Please comment on whether it is good or not.
   b. Do the prediction using test data.
   c. Which variables are significant?
   d. Is there multi-collinearity?
7. How do you improve the accuracy of the model? Write clearly the changes that you will make before re-fitting the model. Fit the final model. **(20 MARKS)**

   Please feel free to have any number of iterations to get to the final answer. Marks are awarded based on the quality of final model you are able to achieve.

8. **Summarize as follows (10 MARKS)**
   a. *Summarize the overall fit of the model and list down the measures to prove that it is a good model*
   b. *Write down a business interpretation/explanation of the model – which variables are affecting the target the most and explain the relationship. Feel free to use charts or graphs to explain.*
   c. *What changes from the base model had the most effect on model performance?*
   d. *What are the key risks to your results and interpretation?*